PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2001-167124
(43)Date of publication of application : 22.06.2001

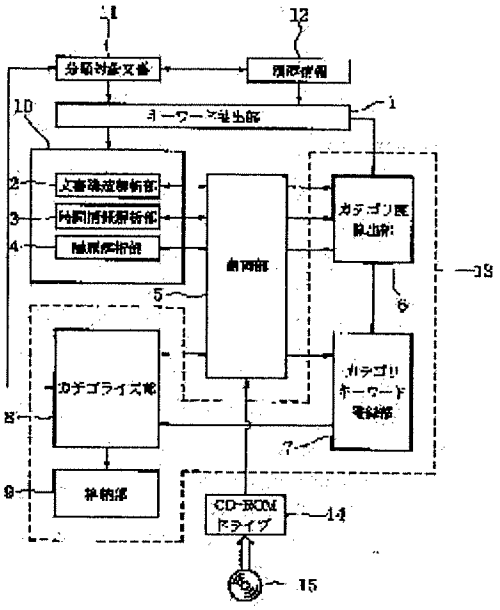
(51)Int.Cl. G06F 17/30

(21)Application number : 11-353556
(22)Date of filing : 13.12.1999

(71)Applicant : SHARP CORP
(72)Inventor : IWATA NOBUYUKI

(54) DOCUMENT CLASSIFICATION DEVICE AND RECORDING MEDIUM RECORDING DOCUMENT CLASSIFICTION PROGRAM

(57)Abstract:
PROBLEM TO BE SOLVED: To automatically classify massive documents to a category which has a current topic or is adjusted to the taste of a user without permitting the user to consciously decide the category for classification.
SOLUTION: A document classification device is provided with a storage part (not illustrated) storing the document being the object of classification and the document of history information 12, a keyword extraction part 1 extracting a keyword from the document stored in the storage part, an analysis part 10 calculating weight by setting the keyword to be the classification destination of the document based on the significance of the keyword extracted by the keyword extraction part 1 and at least, the preservation date/time or the preservation place of the document comprising the keyword and a classification part 13 classifying the document based on the weights calculated by the analysis part 10.



(19)日本国特許庁（ＪＰ）　(12)**公開特許公報**（Ａ）　(11)特許出願公開番号
特開2001－167124
(P2001－167124A)
(43)公開日　平成13年６月22日(2001.6.22)

(51)Int.Cl.⁷	識別記号	ＦＩ		テーマコード（参考）
Ｇ０６Ｆ　17/30		Ｇ０６Ｆ	15/401	３１０Ｄ　５Ｂ０７５
			15/40	３７０Ａ
			15/401	３１０Ａ
			15/403	３４０Ａ

審査請求　未請求　請求項の数６　ＯＬ　（全９頁）

(21)出願番号　特願平11－353556

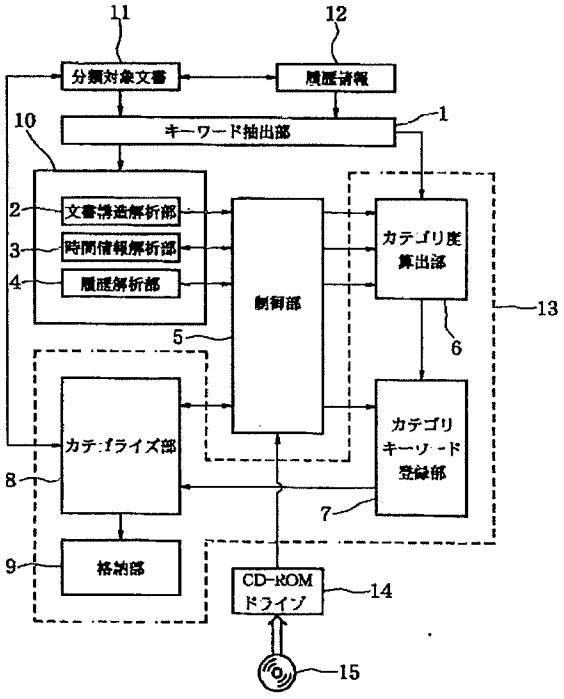
(22)出願日　平成11年12月13日(1999.12.13)

(71)出願人　000005049
シャープ株式会社
大阪府大阪市阿倍野区長池町22番22号
(72)発明者　岩田　展幸
大阪府大阪市阿倍野区長池町22番22号　シャープ株式会社内
(74)代理人　100091096
弁理士　平木　祐輔

Ｆターム(参考)　5B075 ND03 NK32 NR03 NR12 PR08 UU06

(54)【発明の名称】　文書分類装置及び文書分類プログラムを記録した記録媒体

(57)【要約】
【課題】　ユーザが意識的に分類のためのカテゴリを決定することなく、話題性、又はユーザの嗜好に合ったカテゴリへ大量の文書を自動分類する。
【解決手段】　本発明の文書分類装置は、分類対象文書１１及び履歴情報１２の文書を記憶する記憶部（図示せず）と、記憶部に記憶されている文書からキーワードを抽出するキーワード抽出部１と、キーワード抽出部１により抽出されるキーワードの重要度、該キーワードを含む文書の保存日時又は保存場所の少なくとも一つに基づいて、該キーワードを上記文書の分類先にする重みを算出する解析部１０と、解析部１０により算出される各重みに基づいて、文書を分類する分類部１３と、を備える。



【特許請求の範囲】
【請求項１】　文書を記憶する記憶手段と、前記記憶手段に記憶されている前記文書からキーワードを抽出する抽出手段と、前記抽出手段により抽出されるキーワードの重要度、前記キーワードを含む文書の保存日時又は保存場所の少なくとも一つに基づいて、前記キーワードを前記文書の分類先にする重みを算出する重み算出手段と、前記重み算出手段により算出される前記重みに基づいて、前記文書を分類する分類手段と、を備えることを特徴とする文書分類装置。
【請求項２】　前記重み算出手段は、前記キーワードが前記文書の見出し、表題、図題、ハイパーリンク又は強調表示、又は固有名詞の少なくとも一つに使用されている場合に前記重みを変える、ことを特徴とする請求項１記載の文書分類装置。
【請求項３】　前記重み算出手段は、前記キーワードを含む前記文書の前記保存日時から前記キーワードが出現する日時を算出し、該算出日時と所定の基準日時との比較に基づいて前記重みを算出する、ことを特徴とする請求項１記載の文書分類装置。
【請求項４】　前記重み算出手段は、前記キーワードを含む前記文書が所定のブラウザのキャッシュ、該ブラウザのお気に入りファイル又はブックマークのリンク先、検索時の文字入力列、ユーザが行った階層分類構造、又は自動分類された階層構造の少なくとも一つである場合に前記重みを変える、ことを特徴とする請求項１記載の文書分類装置。
【請求項５】　前記重み算出手段により算出される前記重みに対する比重を制御する制御手段を更に備え、前記分類手段は、前記制御手段からの比重と前記算出手段により算出される前記重みとに基づいて、前記文書を分類する、ことを特徴とする請求項１記載の文書分類装置。
【請求項６】　コンピュータを、文書を記憶する記憶手段と、前記記憶手段に記憶されている前記文書からキーワードを抽出する抽出手段と、前記抽出手段により抽出されるキーワードの重要度、前記キーワードを含む文書の保存日時又は保存場所の少なくとも一つに基づいて、前記キーワードを前記文書の分類先にする重みを算出する重み算出手段と、前記重み算出手段により算出される前記重みに基づいて、前記文書を分類する分類手段と、を備える文書分類装置として機能させるためのプログラムを記録したことを特徴とするコンピュータ読み取り可能な記録媒体。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】本発明は文書分類装置に関し、特に、情報を蓄積するシステムにおいて文書のグループから文書を分類する文書分類装置及び文書分類プログラムを記録した記録媒体に関する。
【０００２】
【従来の技術】インターネット等の情報通信網の普及に伴い、情報提供者等から必要に応じて取得した文書情報を分類する機会が増えている。
【０００３】本発明に関連する公知技術として、特開平６－３４８７５５号公報、及び特開平１１－１５８４８号公報に開示されている文書分類方法がある。上記各公報に記載されている文書分類方法は、分類済みの文書群から各分野に固有の単語（キーワード）を抽出し、分類対象の文書におけるキーワードの出現頻度に基づいて、分類対象文書の分類先を決定するものである。
【０００４】また、特開平７－１１４５７２号公報に開示されている文書分類方法は、分類対象文書に含まれる単語の特徴を表現する特徴ベクトルから文書の特徴を表現する文書ベクトルを生成し、文書ベクトル間の類似度を利用して分類するものである。
【０００５】
【発明が解決しようとする課題】上記各公報にみられるように、従来の文書分類システムでは、分類対象文書の分類先を決定する要素として、キーワードの文書構造又はキーワードの出現頻度等、文書自身が有する情報を利用しており、話題性のあるカテゴリや、ユーザの嗜好に合ったカテゴリに分類対象の文書を自動分類するものではない。
【０００６】本発明の目的は、話題性又はユーザの嗜好に適したカテゴリに分類対象の文書を自動的に分類することができる文書分類装置、及び文書分類プログラムを記録した記録媒体を提供することにある。
【０００７】
【課題を解決するための手段】上記目的を達成するために本発明の文書分類装置は、文書を記憶する記憶手段と、前記記憶手段に記憶されている前記文書からキーワードを抽出する抽出手段と、前記抽出手段により抽出されるキーワードの重要度、前記キーワードを含む文書の保存日時又は保存場所の少なくとも一つに基づいて、前記キーワードを前記文書の分類先にする重みを算出する重み算出手段と、前記重み算出手段により算出される前記重みに基づいて、前記文書を分類する分類手段と、を備えるものである。
【０００８】また、前記重み算出手段は、前記キーワードが前記文書の見出し、表題、図題、ハイパーリンク又は強調表示、又は固有名詞の少なくとも一つに使用されている場合に前記重みを変えるものであることで、キーワードの重要度を重視して文書を分類できる。
【０００９】また、前記重み算出手段は、前記キーワードを含む前記文書の前記保存日時から前記キーワードが出現する日時を算出し、該算出日時と所定の基準日時との比較に基づいて前記重みを算出するものであることで、キーワードの出現日時等が時間情報として付与され、話題性のあるカテゴリを検出することが可能になる。

【００１０】また、前記重み算出手段は、前記キーワードを含む前記文書が所定のブラウザのキャッシュ、該ブラウザのお気に入りファイル又はブックマークのリンク先、検索時の文字入力列、ユーザが行った階層分類構造、又は自動分類された階層構造の少なくとも一つである場合に前記重みを変えるものであることで、操作履歴情報等が付与され、嗜好性のあるカテゴリを検出することが可能になる。

【００１１】また、さらに、前記重み算出手段により算出される前記重みに対する比重を制御する制御手段を更に備え、前記分類手段は、前記制御手段からの比重と前記算出手段により算出される前記重みとに基づいて、前記文書を分類することにより、ユーザニーズをより考慮して文書を分類できる。

【００１２】他の観点において本発明は、コンピュータを、文書を記憶する記憶手段と、前記記憶手段に記憶されている前記文書からキーワードを抽出する抽出手段と、前記抽出手段により抽出されるキーワードの重要度、前記キーワードを含む文書の保存日時又は保存場所の少なくとも一つに基づいて、前記キーワードを前記文書の分類先にする重みを算出する重み算出手段と、前記重み算出手段により算出される前記重みに基づいて、前記文書を分類する分類手段と、を備える文書分類装置として機能させるためのプログラムを記録したことを特徴とするコンピュータ読み取り可能な記録媒体である。

【００１３】

【発明の実施の形態】本発明の実施の形態を添付図面と対応して以下に詳細に説明する。図１は、本発明の実施の形態による文書分類装置の構成を示すブロック図である。本文書分類装置は、キーワード抽出部（抽出手段）１、制御部（制御手段）５、解析部（重み算出手段）１０、分類部（分類手段）１３及びＣＤ－ＲＯＭドライブ１４を備える。

【００１４】また、解析部１０は、文書構造解析部２、時間情報解析部３及び履歴解析部４を有し、分類部１３は、カテゴリ度算出部６、カテゴリキーワード登録部７、カテゴライズ部８及び格納部９を有する。また、ＣＤ－ＲＯＭドライブ１４は、制御部５の指示に基づいてＣＤ－ＲＯＭ１５に格納されているプログラムを読み出す。

【００１５】上記文書構造解析部２、時間情報解析部３及び履歴解析部４は、全体として、キーワードの重要度、該キーワードを含む文書の保存日時又は保存場所に基づいて該キーワードを文書の分類先にする重みを算出する解析部１０を構成している。

【００１６】分類対象文書１１は、本文書分類装置において分類の対象となる文書である。分類対象文書１１は、例えばインターネット等からダウンロードされ一時的にＨＤＤ等の記録装置に格納されている文書でも良いし、履歴情報１２である文書の集合における要素であっても良い。分類対象文書１１及び履歴情報１２は、各文書の保存日時、保存場所等の情報が認識できる状態で、ＨＤＤ等の記録装置（記憶手段）に格納されている。

【００１７】履歴情報１２は、本文書分類装置により管理される文書の集合である。履歴情報１２は、例えばWorld Wide Webブラウザ（以下、Webブラウザと呼ぶ）のキャッシュの文書、Webブラウザのお気に入りファイル又はブックマークのリンク先の文書、検索操作時の入力文字列、または、既にユーザが行った階層分類構造又は自動分類された階層構造等である。

【００１８】キーワード抽出部１は、分類対象文書１１及び履歴情報１２の文書からキーワードを抽出する。分類対象文書１１及び履歴情報１２の文書におけるある文字列が、カテゴリキーワード登録部７のキーワード辞書（図示せず）に予め登録されているカテゴリキーワードと一致する場合に、該文字列をキーワードとして抽出する。

【００１９】解析部１０において、文書構造解析部２は、キーワード抽出部１により抽出されたキーワードの文書構造上の重みを構造重みとして算出する。該キーワードの構造重みが大きいほど、文書中で重要なキーワードとなる。

【００２０】例えば、構造重みが大きくなるのは、分類対象文書１１又は履歴情報１２の文書における見出し、表題、図題に使用されているキーワード、及びハイパーリンク等で使用されているキーワードである。また、強調表示されているキーワード、他のキーワードと比較して大きなフォントで表示されているキーワード、及び固有名詞で用いられているキーワードの構造重みも大きくなる。この構造重みは、制御部５からの指示で変化させる。

【００２１】また、時間情報解析部３は、キーワード抽出部１により抽出されたキーワードの時間軸上の重みを時間重みとして算出する。分類対象文書１１から抽出されたキーワードを含む履歴情報１２の文書が有する更新日時、保存日時等の情報から該キーワードの出現頻度の最も高い日時を算出する。この場合、該キーワードの出現頻度が所定の頻度を超える日時をキーワード出現日時として算出するようにしても良い。

【００２２】例えば、算出日時が基準日時に近いキーワードほど時間重みは大きくなり、算出日時が基準日時に遠いキーワードほど時間重みは小さくなる。この時間重みは、制御部５からの指示で変化させる。また、履歴解析部４は、キーワード抽出部１により抽出されたキーワードが履歴情報１２の文書の中で出現する頻度を履歴重みとして算出する。

【００２３】例えば、Webブラウザで最近閲覧してキャッシュ中に記憶されているページに使用されている場合、又は検索操作した際に使用された文字列等に使用されている場合に履歴重みは大きくなる。この履歴重み

は、制御部5からの指示で変化させる。

【0024】このように、本実施の形態では、文書構造解析部2、時間情報解析部3及び履歴解析部4が設けられており、分類対象文書11から抽出したキーワードが該分類対象文書11の分類先として適切であるか否かを示す指標として、文書構造解析部2により重要性のあるカテゴリを検出することが可能になり、時間情報解析部3によりキーワードの出現日時を時間情報として付与して話題性のあるカテゴリを検出することが可能となるとともに、履歴解析部4により嗜好性のあるカテゴリを検出することが可能となる。

【0025】制御部5は、以下に示す4つの制御を主に行う。第1の制御として、文書構造解析部2、時間情報解析部3、履歴解析部4にて算出される重みの基準値を設定する。文書構造解析における基準値には、見出し、表題、図題、ハイパーリンク、強調表示、及びフォントを要素とする基準ベクトルを設定する。この基準ベクトルに対してどの要素に比重を置いて重みを算出するのかを決定する。また、時間情報解析における基準値には、時間軸上の値を設定する。設定した基準値に近いキーワードの時間重みは大きくなる。

【0026】また、履歴情報解析における基準値には、履歴情報12の文書を要素とする基準ベクトルを設定する。この基準ベクトルに対してどの要素に比重を置いて重みを算出するのかを決定する。例えば「検索操作時の入力文字列」という要素を大きくした基準ベクトルを基準値として設定した場合には、「検索操作時の入力文字列」に使用されているキーワードに大きな比重が置かれ、履歴重みが算出される。

【0027】第2の制御として、文書構造解析部2、時間情報解析部3、履歴解析部4で算出した重みから、カテゴリ度算出部6においてカテゴリ度を導く際にそれぞれの重みの比重を制御する。

【0028】第3の制御として、カテゴリキーワードの数を制御する。分類対象文書11の数が多い場合にはカテゴリキーワードとして登録するキーワード数を増加させ、逆に分類対象文書11の数が少ない場合にはカテゴリキーワードとして登録するキーワード数を減少させる。

【0029】第4の制御として、分類先の文書数の最大値及び最小値を制御する。分類先の文書数が最大値を超えた場合には、新たにカテゴリキーワードを登録して再分類する。また、分類先の文書数が最小値に満たない場合には、そのカテゴリを分類先とせず他のカテゴリに分類する。

【0030】カテゴリ度算出部6は、文書構造解析部2、時間情報解析部3及び履歴解析部4からキーワード毎にカテゴリベクトルを導き、カテゴリベクトルの大きさからカテゴリ度を算出する。カテゴリベクトルは、構造重み、時間重み及び履歴重みの3要素からなり、各重みに対する比重は制御部5により与えられる。

【0031】上述のように、カテゴリベクトルを導出する際には、各重みに対する比重を変更できる。文書構造上の重みの比重を大きくした場合には、キーワードの重要性を重視したカテゴリベクトルが導出される。文書の保存日時の情報に基づく話題的要素である時間重みの比重を大きくした場合には、話題性を重視したカテゴリベクトルが導出される。文書の格納場所の情報を利用してユーザの操作履歴の情報による嗜好的要素である履歴重みの比重を大きくした場合には、ユーザの嗜好を重視したカテゴリベクトルが導出される。

【0032】カテゴリキーワード登録部7は、カテゴリ度算出部6で算出されたカテゴリ度に基づいて、カテゴリキーワードを登録する。例えば、分類対象文書11及び履歴情報12の文書から抽出したキーワードをカテゴリ度の高いものから順に、制御部5から与えされたカテゴリキーワード数の上限まで登録する。

【0033】カテゴライズ部8は、カテゴリキーワード登録部7により登録されたカテゴリキーワードの中から、カテゴリ度が最も高いカテゴリキーワードを分類対象文書11の分類先としてカテゴライズする。格納部9は、カテゴライズ部8からの分類先の指示に基づいて分類対象文書11を格納する。

【0034】次に、図2から図6に示すフローチャートを参照して、本実施の形態による文書分類装置の動作を説明する。図2は、本実施の形態による文書分類装置の文書分類処理を説明するフローチャートである。はじめに、ステップS11では、分類対象文書11から抽出したキーワードの中からカテゴリキーワードを登録し、次いで、ステップS12で、分類対象文書11をカテゴライズする。

【0035】ステップS13では、分類先の文書数が、設定されている基準最小値以下であるか否かを判別する。ここで、分類先の文書数が基準最小値以下である場合には、分類先として設定されているカテゴリキーワードはカテゴリとして相応しくないと判断して削除し、ステップS12に戻りカテゴライズ処理を再度実行する。また、分類先の文書数が基準最小値より大きい場合には、ステップS14に進む。

【0036】ステップS14では、分類先の文書数が設定されている基準最大値以上であるか否かを判別する。ここで、分類先の文書数が基準最大値以上である場合には、より小さい単位のカテゴリに分類可能であると判断してカテゴリキーワードを新しく追加登録し、ステップS12に戻りカテゴライズ処理を再度実行する。また、分類先の文書数が基準最大値より小さい場合には、ステップS15に進み分類対象文書11を分類先に各々格納する。

【0037】図3は、本実施の形態による文書分類装置のカテゴリキーワード登録処理を説明するフローチャー

トであり、図2のステップS11の処理に対応する。はじめに、ステップS21では、分類対象文書11及び履歴情報12の文書からキーワードを抽出し、次いで、ステップS22で、抽出されたキーワード毎にカテゴリ度を算出する。

【0038】次いで、ステップS23で、算出されたカテゴリ度と基準値とを比較する。ここで、算出されたカテゴリ度が基準値以上の場合には、ステップS24に進みカテゴリキーワードとして登録する。また、算出されたカテゴリ度が基準値以下の場合にはステップS25に進む。

【0039】ステップS25では、全てのキーワードを検索したか否かを判別し、全てのキーワードが検索されたと判断された場合には、カテゴリキーワードの登録処理を完了する。また、全てのキーワードが検索されていない場合には、ステップS21に戻り次のキーワードの処理に移る。

【0040】図4は、本実施の形態による文書分類装置のカテゴリ度算出処理を説明するフローチャートであり、図3のステップS22の処理に対応する。はじめに、ステップS31では、文書構造に基づく構造重みを算出する。文書構造解析部2は、キーワード抽出部1で抽出されたキーワードのうち、分類対象文書11に含まれるキーワードの文書構造を調べ、キーワード毎に時間重みを算出する。上述のように文書構造は、分類対象文書11と履歴情報12の文書において、該キーワードが使用されている場所又は文字サイズ等を示す。

【0041】構造重みの基準値には、見出し、表題、図題、ハイパーリンク、強調表示、及びフォントを要素とする基準ベクトルが設定される。基準ベクトルに基づいた配分で重み付けを行う。例えば、ハイパーリンクに比重を置いた基準ベクトルに基づいた重み付けでは、ハイパーリンクに使用されているキーワードの構造重みは大きくなる。この基準ベクトルは予め設定されているが、ユーザが適宜変更することができる。

【0042】次いで、ステップS32で、日時情報に基づく時間重みを算出する。時間情報解析部3は、キーワード抽出部1で抽出されたキーワードのうち、分類対象文書11に含まれるキーワードを含む文書が有する日時情報を調べ、キーワード毎に時間重みを算出する。上述のように日時情報は、該キーワードが出現した日時を示す。ただし、複数の文書で該キーワードが出現した場合には、各文書内での出現回数と各文書の保存日時から算出した値を日時情報とする。時間重み算出における基準値には、時間軸上の値を設定する。設定された基準値に近いキーワードの時間重みは大きくなる。

【0043】ステップS33では、履歴情報12に基づく履歴重みを算出する。履歴解析部4は、キーワード抽出部1で抽出されたキーワードのうち、分類対象文書11に含まれるキーワードを含む文書の保存場所を調べ、キーワード毎に履歴重みを算出する。

【0044】履歴重みの基準値には、履歴情報12の文書を要素とする基準ベクトルを設定する。基準ベクトルに基づいてどの要素に比重を置いて重みを算出するのかを決定する。例えば、「検索操作時の入力文字列」という要素を大きくしたベクトルを基準値として設定すると、「検索操作時の入力文字列」に使用されているキーワードに大きな比重をおかれ、履歴重みが算出される。

【0045】ステップS34では、キーワード毎に算出された構造重み、時間重み及び履歴重みを要素とするカテゴリベクトルを導出する。導出されたカテゴリベクトルは、制御部5からの各重みに対する比重を考慮して各重みの大きさが変更される。次いで、ステップS35で、導出されたカテゴリベクトルの大きさからカテゴリ度を算出する。

【0046】図5は、本実施の形態による文書分類装置のカテゴライズ処理を示すフローチャートであり、図2のステップS12の処理に対応する。はじめに、ステップS41では、分類対象文書11に対して登録されているカテゴリキーワードを検索し、次いで、ステップS42で、分類対象文書11からカテゴリキーワードが見つかったか否かを判断する。

【0047】ステップS42において、分類対象文書11からカテゴリキーワードが見つかった場合には、ステップS43に進む。ステップS43では、カテゴリキーワードの中で最もカテゴリ度の高いものを検索し、最もカテゴリ度が高いカテゴリキーワードをカテゴリに分類する。

【0048】また、該分類対象文書11からカテゴリキーワードが見つからなかった場合には、ステップS44に進み類似カテゴリ検索を行う。ステップS44では、該分類対象文書11が最も類似するカテゴリキーワードを分類先とする。

【0049】図6は、本実施の形態による文書分類装置の類似カテゴリ検索処理を示すフローチャートであり、図5のステップS44の処理に対応する。分類対象文書11中にカテゴリキーワードとなるキーワードが存在しないので、まずステップS51で、該分類対象文書11から抽出された各キーワードに対応するカテゴリベクトルの平均を算出し、算出した平均を該分類対象文書自身のカテゴリベクトルとする。

【0050】次いで、ステップS52で、ステップS51で算出した文書のカテゴリベクトルと、カテゴリキーワードのカテゴリベクトルとの類似度を算出する。次いで、ステップS53で、算出された類似度が最大値を超えたか否かを判別する。

【0051】ステップS53において、算出された類似度が最大値を超える場合には、ステップS54及びステップS55で、類似度の最大値及び類似カテゴリを更新する。また、算出された類似度が最大値を超えない場合

には、ステップS56に移る。

【0052】ステップS56では、文書のカテゴリベクトルと、全てのカテゴリキーワードのカテゴリベクトルとの比較が終了したか否かを判別する。ここで、全てのカテゴリキーワードのカテゴリベクトルとの比較が終了した場合には類似カテゴリ検索処理を終了する。また、全てのカテゴリキーワードのカテゴリベクトルとの比較が終了していないと判断された場合には、ステップS52の処理に戻り、次のカテゴリキーワードの類似カテゴリ検索を行う。以上により、最も類似度が大きいカテゴリキーワードが該分類対象文書11の分類先となる。

【0053】以上説明したように、本実施の形態の文書分類装置は、管理可能な文書の集合からキーワードを抽出するキーワード抽出部1と、分類対象文書11から抽出したキーワードに対して、文書中の構造に基づく構造重みを算出する文書構造解析部2と、文書の作成日時情報からキーワード出現の時間的な位置付けを時間重みとして算出する時間情報解析部3と、ユーザの操作履歴情報を履歴重みとして算出する履歴解析部4と、上記構造重みと、時間重み及び履歴重みとからキーワードのカテゴリベクトルを導出し、該ベクトルの大きさをカテゴリ度とするカテゴリ度算出部6と、該カテゴリ度を有するキーワードからカテゴリとなるカテゴリキーワードを選出し登録するカテゴリキーワード登録部7と、分類対象文書11の分類先を判別するカテゴライズ部8と、該分類先に分類対象文書11を格納する格納部9と、分類先集合内の文書数により分類を制御する制御部5を備え、分類対象文書11から抽出したキーワードのみでなく、文書作成日時等の時間情報や蓄積しているユーザの操作履歴情報を利用して文書分類先を決定するように構成したので、話題性を加味したカテゴライズとユーザの嗜好に合ったカテゴライズを実現することができる。

【0054】なお、本実施の形態では、解析部10により構造重み、時間重み及び履歴重みを各々算出し、分類部13により制御部5からの重みの比重の指示に基づいて各重みを要素とするカテゴリベクトルを導出している。本発明は上記実施の形態に限定されず、制御部5の比重の指示による特別なケースとして、キーワードの重要度（構造重み）、話題性（時間重み）又はユーザの嗜好（履歴重み）の少なくとも一つを重視し、重視した重みのみを要素とするカテゴリベクトルを導出して文書分類処理を行う構成であれば良い。

【0055】また、本実施の形態では、分類部13をカテゴリ度算出部6、カテゴリキーワード登録部7、カテゴライズ部8及び格納部9に分けて示しているが本発明を限定するものではなく、解析部10により算出される各重みに基づいて、分類対象文書11を自動的に分類する構成であれば良い。

【0056】上述したように、本発明の文書分類装置は、本文書分類装置を機能させるためのプログラムでも実現される。このプログラムはコンピュータで読み取り可能な記録媒体に格納されている。本発明では、この記録媒体としてROM（図示せず）そのものがプログラムメディアであっても良いし、また、外部記憶装置としてCD-ROMドライブ14等のプログラム読み取り装置が設けられ、そこに記録媒体を挿入することで読み取り可能なCD-ROM15等のプログラムメディアであっても良い。いずれの場合においても、格納されているプログラムは制御部5がアクセスして実行させる構成であっても良いし、プログラムを読み出し、読み出されたプログラムは、図示されていないプログラム記憶エリアにダウンロードされて、そのプログラムが実行される方式であっても良い。このダウンロード用のプログラムは予め本体装置に格納されているものとする。

【0057】ここで上記プログラムメディアは、本体と分離可能に構成される記録媒体であり、磁気テープやカセットテープ等のテープ系、フロッピーディスクやハードディスク等の磁気ディスクやCD-ROM／MO／MD／DVD等の光ディスクのディスク系、ICカード（メモリカードを含む）／光カード等のカード系、あるいはマスクROM、EPROM、EEPROM、フラッシュROM等による半導体メモリを含めた固定的にプログラムを担持する媒体であっても良い。

【0058】さらに、送受信手段（図示せず）を介して通信ネットワーク（図示せず）からプログラムをダウンロードするように、流動的にプログラムを担持する媒体であっても良い。なお、このように通信ネットワークからプログラムをダウンロードする場合には、そのダウンロード用プログラムは予め装置本体に格納しておくか、あるいは別な記録媒体からインストールされるものであっても良い。なお、記録媒体に格納されている内容としてはプログラムに限定されず、データであっても良い。

【0059】

【発明の効果】本発明によれば、キーワード自身の重要度、話題性又はユーザの嗜好に合ったカテゴリを抽出して、分類対象文書を自動的に分類できる。また、キーワード自身の重要度、話題性及びユーザの嗜好の各々に対応する重みの比重を自由に変更することにより、ユーザニーズに適したシステムを実現できる。

【図面の簡単な説明】

【図1】本発明の実施の形態による文書分類装置の構成を説明するブロック図である。

【図2】本発明の実施の形態による文書分類装置の文書分類処理を説明するフローチャートである。

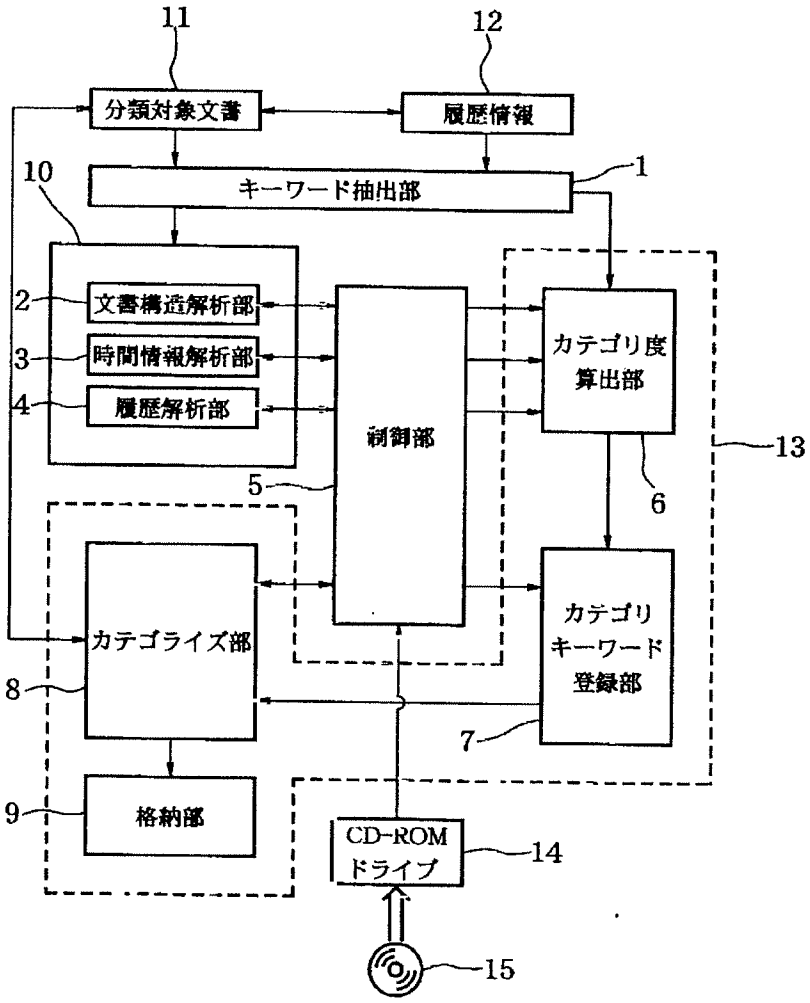
【図3】本発明の実施の形態による文書分類装置のカテゴリキーワード登録処理を説明するフローチャートである。

【図4】本発明の実施の形態による文書分類装置のカテゴリ度算出処理を説明するフローチャートである。

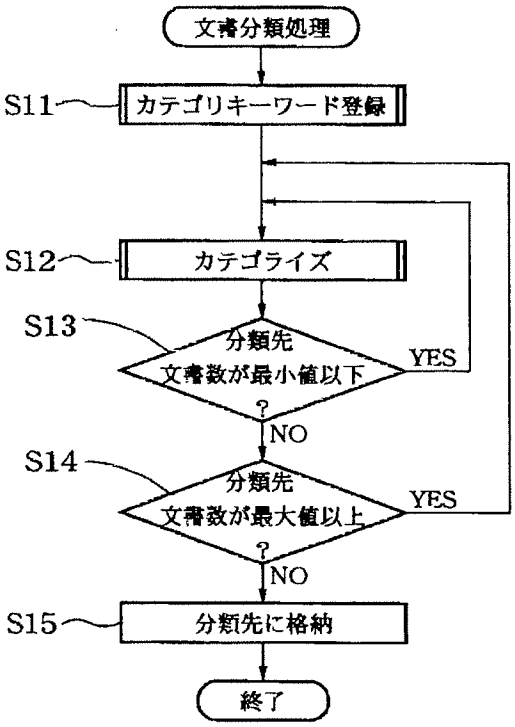
【図5】本発明の実施の形態による文書分類装置の文書

のカテゴライズ処理を説明するフローチャートである。
【図６】本発明の実施の形態による文書分類装置の類似カテゴリ検索処理を説明するフローチャートである。
【符号の説明】
１：キーワード抽出部（抽出手段）
２：文書構造解析部
３：時間情報解析部
４：履歴解析部
５：制御部（制御手段）
６：カテゴリ度算出部
７：カテゴリキーワード登録部
８：カテゴライズ部
９：格納部
１０：解析部（重み算出手段）
１１：分類対象文書
１２：履歴情報
１３：分類部（分類手段）
１４：ＣＤ－ＲＯＭドライブ
１５：ＣＤ－ＲＯＭ（記録媒体）

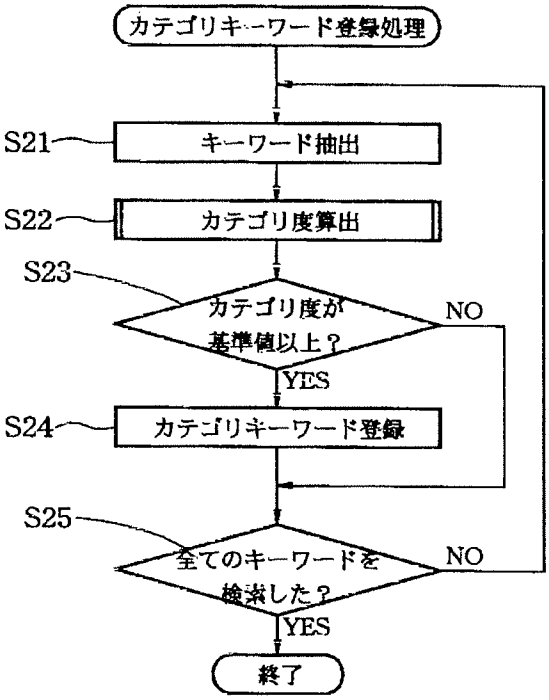
【図１】



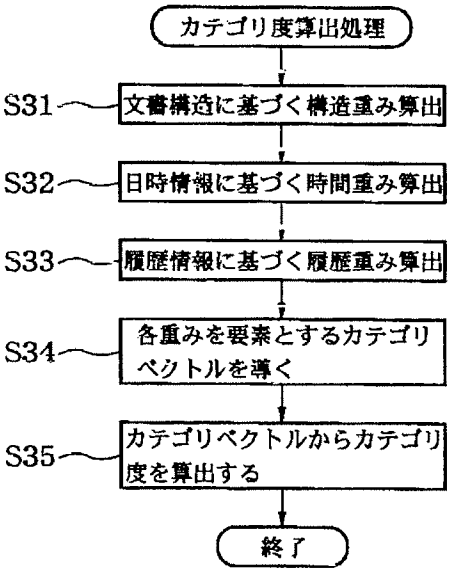
【図2】


文書分類処理
S11
カテゴリキーワード登録
S12
カテゴライズ
S13
分類先
文書数が最小値以下
？
YES
NO
S14
分類先
文書数が最大値以上
？
YES
NO
S15
分類先に格納
終了

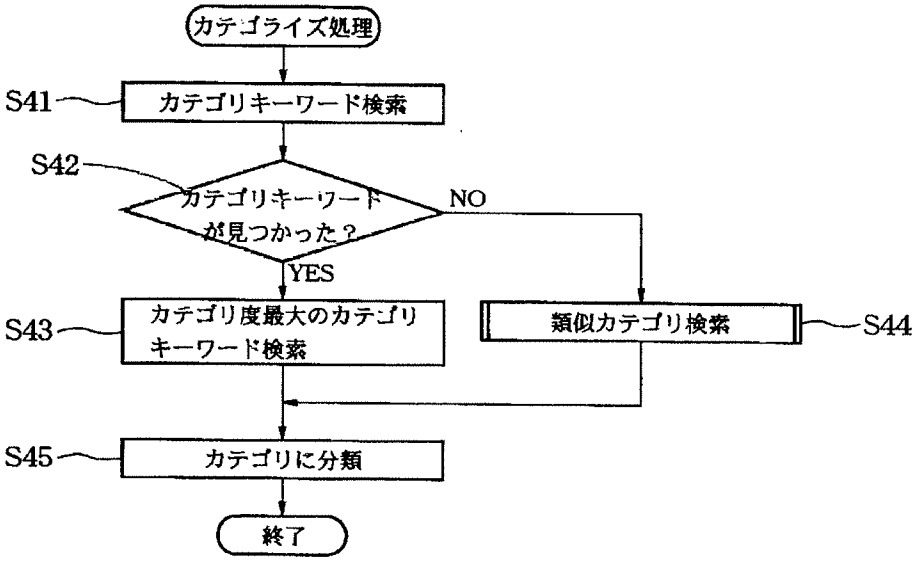
【図3】


カテゴリキーワード登録処理
S21
キーワード抽出
S22
カテゴリ度算出
S23
カテゴリ度が
基準値以上？
NO
YES
S24
カテゴリキーワード登録
S25
全てのキーワードを
検索した？
NO
YES
終了

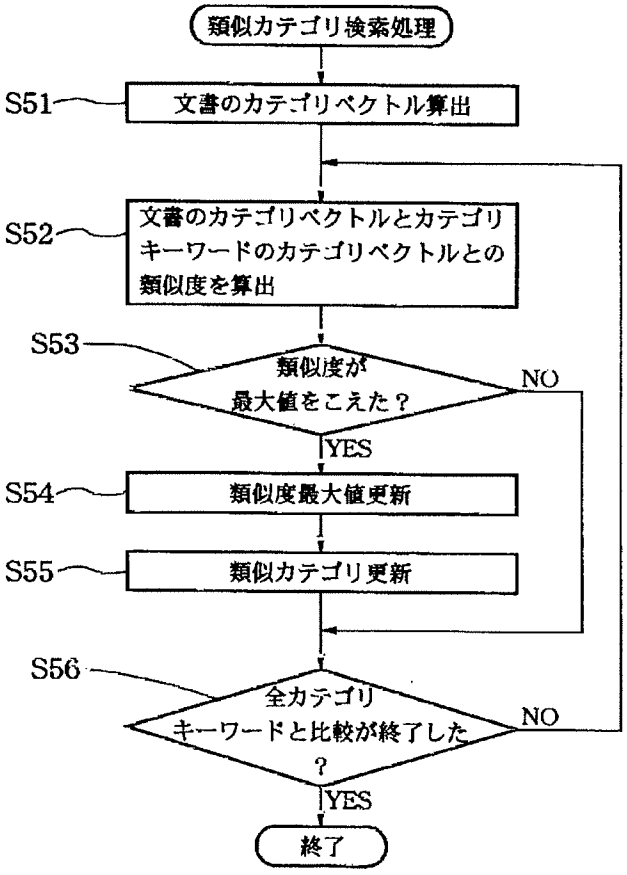
【図4】


カテゴリ度算出処理
S31
文書構造に基づく構造重み算出
S32
日時情報に基づく時間重み算出
S33
履歴情報に基づく履歴重み算出
S34
各重みを要素とするカテゴリ
ベクトルを導く
S35
カテゴリベクトルからカテゴリ
度を算出する
終了

【図5】


カテゴライズ処理
S41
カテゴリキーワード検索
S42
カテゴリキーワードが見つかった？
NO
YES
S43
カテゴリ度最大のカテゴリキーワード検索
類似カテゴリ検索
S44
S45
カテゴリに分類
終了

【図6】


類似カテゴリ検索処理
S51
文書のカテゴリベクトル算出
S52
文書のカテゴリベクトルとカテゴリキーワードのカテゴリベクトルとの類似度を算出
S53
類似度が最大値をこえた？
NO
YES
S54
類似度最大値更新
S55
類似カテゴリ更新
S56
全カテゴリキーワードと比較が終了した？
NO
YES
終了